A Representation Theorem for Guilt Aversion*

Martin Kaae Jensen[†a] and **Maria Kozlovskaya**[a,b]

[a]Department of Economics, University of Leicester, Leicester, LE1 7RH, UK

[b]The Business School, University of Huddersfield, West Yorkshire, HD1 3DH, UK

31st January 2016

Abstract

Guilt aversion has been shown to play an important role in economic decision-making. In this paper, we take an axiomatic approach to guilt by deducing a utility representation from a list of easily interpretable assumptions on an agent's preferences. It turns out that our logarithmic representation can mitigate the problem of multiplicity of equilibria to which psychological games are prone. We apply the model in three well-known games and show that its predictions are consistent with experimental observations.

Keywords: Guilt aversion, utility representation, psychological games.

JEL classification: C72, C91, D03.

*We would like to thank Kirill Borusyak, Subir Bose, Martin Dufwenberg, David Rojo Arjona, Rajiv Sarin, conference participants at the University of Ghent and the University of Manchester, as well as the editor William S. Neilson and two anonymous referees for their helpful comments and constructive suggestions. Maria Kozlovskaya gratefully acknowledges financial support from the ESRC.

[†]Corresponding author. Tel: +44 116 252 2901

E-mail addresses: mj182@leicester.ac.uk (M.K.Jensen), m.kozlovskaya@hud.ac.uk (M.Kozlovskaya)

1 Introduction

Guilt is the experience of discomfort that follows when we violate a personal or social standard. If an action raises income but disappoints our own or other individuals' expectations of us, it may trigger our guilty conscience. Any individual who is sufficiently averse to this discomfort may therefore refrain from taking the action in the first place. Guilt aversion is able to explain a vast array of behaviors, including cooperation (Miettinen and Suetens, 2008), altruism (Andreoni and Rao, 2011), conformism (Khalmetski, 2015), group favoritism (Güth et al., 2009) and reciprocity (Chang et al., 2011), and economic experiments indicate that it is indeed an important determinant in a variety of different situations (Ketelaar and Au, 2003; Charness and Dufwenberg, 2006; Hopfensitz and Reuben, 2009; Geng et al., 2011; Battigalli et al., 2013). More recently, guilt averse behavior has also made its way into macroeconomic modeling. Thus Ahrens and Snower (2014) incorporates guilt and envy into a dynamic stochastic equilibrium model and shows that when these emotions are experienced by workers, a Phillips curve relationship between inflation and output can be generated.

A popular way to model emotions, including guilt, is to include them as inputs in agents' utility functions. Particularly important for experimental work are linear utility representations with money and guilt as the inputs (Battigalli and Dufwenberg, 2007; López-Pérez, 2010; Chang et al., 2011; Battigalli et al., 2013; Miettinen, 2013; Khalmetski et al., 2015).[1] This paper's main theoretical objective is to axiomatize utility representations of guilt-averse preferences. Specifically, axioms are presented that are necessary and sufficient for (i) the linear representation mentioned a moment ago, (ii) a representation that is logarithmic in money and linear in guilt, and (iii) a general additively separable utility representation of money and guilt. Call the sacrifice ratio between money and guilt ("how much money an agent is willing to pay to lower guilt by one unit") the *price of a clear conscience* (PCC). For well-behaved preferences we find that (i) obtains if and only if the PCC is constant for all money-guilt combinations; (ii) holds if and only if for any two levels of income the relative PCC equals the relative income, and (iii) derives whenever a suitably redefined "double cancellation condition" (Debreu, 1960) is satisfied.

By tracing specific utility representations to the level of preferences, we are able to shed light on the deeper psychological conditions that they entail vis-a-vis the previously mentioned personal or social standards. In doing so, we quickly end up concluding that the assumptions about an agent's moral compass embodied in (i) are problematic. While (iii) is not subject to this critique, it has — as will become clear from the following discussion — too many degrees of freedom to provide a useful alternative in strategic settings. This motivates our introduction of (ii) as the simplest realistic alternative — and it is important to stress, this is *not* an ad hoc alternative but one grounded in moral/psychological considerations. With this in hand we then — in what is arguably the paper's main contribution to existing literature — reanalyze a number of famous laboratory games, namely the Dictator game, the Public Good Provision game, and the Prisoners' Dilemma. This exercise provides further support for model (ii), but we postpone the specifics until section 4.

To the best of our knowledge, the only existing paper concerned with the axiomatization of guilt-

[1] Miettinen (2013) considers a linear utility over money and guilt in the main text of his paper. In the appendix he studies a more general function with a weakly convex guilt component, which he adopts for technical convenience but finds difficult to justify. The alternative functional form proposed in this paper also implies convex preferences over money and guilt, but is grounded in deep psychological considerations.

income representations is López-Pérez (2010). López-Pérez (2010) proposes a utility function exhibiting guilt aversion and provides axiomatic foundation for it. The study also features a discussion of the psychological foundation of guilt and shame and links the feeling of guilt to internalization of a social norm. The paper differs from ours in a number of ways, however, most importantly in the definition of guilt. In López-Pérez (2010), guilt is binary (-1 if the social norm was breached and 0 otherwise), and for the value of guilt to be determined, an exogenous social norm must be specified. By contrast, in our setting guilt is a real number with the standard interpretation as the difference between an opponent's actual and expected payoff (see *e.g.* Battigalli and Dufwenberg (2007) as well as the discussion in section 2). Finally, the properties of the preference relation in López-Pérez (2010) depend on what other players do, hence any given representation is only defined within a specific game. By contrast, our preference relation is set on an abstract guilt-money space and thus can be applied to both decision and game theory. An axiomatic approach to the broader field of other-regarding preferences has been pursued by several authors, most notably Neilson (2006) and Sandbu (2008). Both papers axiomatize general function forms: Additively separable reference-dependent utility in the former and CES-utility in the latter. What sets these studies apart from the results of the current paper is our focus on specific functional forms with few enough free parameters to be testable in the laboratory.

The rest of the paper proceeds as follows. Section 2 introduces money-guilt utility functions. Section 3 develops a theory of moral choice and presents our axiomatization results. Section 4 studies the experimental evidence in the three games mentioned above as well as further discussion. The Appendix contains proofs.

2 Existing Literature and the Logarithmic Alternative

The first formal model of guilt aversion was proposed by Battigalli and Dufwenberg (2007). They define guilt as the perceived payoff loss inflicted on another player, *i.e.*, as the difference between an opponent's expected payoff $E(m_j)$ and actual payoff m_j:

$$G(m_j, E(m_j)) = \max\{0, E(m_j) - m_j\}. \tag{1}$$

To be precise, since a player i does not know exactly how much his opponent j expects, $E(m_j)$ is i's belief about j's expectation. That makes guilt, and a guilt-averse agent's utility, a function of second-order beliefs (cf. Geanakoplos et al. (1989), Attanasi and Nagel (2007), Battigalli and Dufwenberg (2009)). Battigalli and Dufwenberg (2007) also propose a utility function over money and guilt (2), which has been extensively used in subsequent theoretical and experimental research.[2]

$$u_i(m_i, G) = m_i - \theta G. \tag{2}$$

Here m_i is the decision-maker's monetary payoff, G is the guilt he experiences, and θ is a guilt sensitivity parameter. A key advantage of such an approach is that it endogenizes the reference point $E(m_j)$ which with a formulation such as (2) is implicitly solved for in equilibrium. A constant marginal rate of

[2]Examples are Battigalli and Dufwenberg (2009), Chang et al. (2011), Battigalli et al. (2013), Khalmetski (2015).

substitution (MRS) between money and guilt arguably has a drawback, however: It can explain nearly any observed behavior. To illustrate with an often studied example, consider the so-called Dictator game, in which one player (the Dictator, hereafter D) decides upon the division of the total endowment T between himself and the other player (the Recipient, hereafter R). His donation to R, m_R, is hence his strategy. In Psychological Nash equilibrium of the game, D's donation will maximize his utility, given his belief about what R expects from him ($E(m_R)$), and this belief will be correct ($E(m_R) = m_R$). It is easy to see that, if D's utility is defined as in (2): $u = T - m_R - \theta \max\{0, E(m_R) - m_R\}$, either the set of equilibria coincides with the strategy set (if $\theta \geq 1$), or giving zero is the only equilibrium (if $\theta < 1$). The second case is falsified by experimental evidence of positive giving in the Dictator game (Engel, 2011). The first case is consistent with experimental data but — and this is our main point — it is unfalsifiable in the sense that it is consistent with *any* set of empirical/experimental observations.

In section 4, we return to the Dictator game as well as two other games that suffer from related difficulties and explain how these shortcomings are overcome if we instead use the following logarithmic specification:

$$u_i(m_i, G) = \log m_i - \theta G. \tag{3}$$

To be sure, it is rather obvious that one way out of the previously described predicament is *not* to assume a constant MRS.[3] At the same time, one cannot pass to arbitrary utility representations, however, including arbitrary additively separable representations, since that merely compounds the explanatory richness. In brief, one must commit to a specific functional form for falsification to be possible. In the next section we will argue from deeper moral axioms that (3) is a proper alternative to linear specifications.

3 Moral Choice and Axiomatization

Consider a decision-maker who has a preference relation $\succeq$ over a two-dimensional choice set $\mathfrak{M} \times \mathfrak{G} \subseteq R^2_+$, where $m \in \mathfrak{M} = R_{++}$ is his strictly positive monetary payoff, and $G \in \mathfrak{G} = R_+$ is the guilt he experiences. $(m_1, G_1) \succeq (m_2, G_2)$ reads "a payoff of m_1 accompanied by guilt of size G_1 is at least as good as a payoff of m_2 accompanied by guilt of size G_2".

We assume throughout that $\succeq$ is complete and transitive (rational), monotone in the sense that money is desirable whereas guilt is undesirable ($m_1 > m_2 \Rightarrow (m_1, G) \succ (m_2, G)$ and $G_1 < G_2 \Rightarrow (m, G_1) \succ (m, G_2)$), and continuous (for all (m_1, G_1), the lower and the upper contour sets, $\{(m, G) : (m, G) \succeq (m_1, G_1)\}$ and $\{(m, G) : (m, G) \preceq (m_1, G_1)\}$ are closed). These assumptions are of course completely standard.

The previous assumptions together with convexity of the choice set $\mathfrak{M} \times \mathfrak{G}$ imply the existence of a continuous utility representation, *i.e.*, a continuous function $u : \mathfrak{M} \times \mathfrak{G} \to \mathbb{R}$ so that $u(m_1, G_1) \geq u(m_2, G_2) \Leftrightarrow (m_1, G_1) \succeq (m_2, G_2)$ (Debreu, 1954). A simple adaption of another contribution by Debreu, immediately provides us with necessary and sufficient conditions for u to be *additively separable*, *i.e.*, for u

[3] Note that a decreasing MRS between money and guilt implies strictly convex preferences over the player's own and his opponent's income – an idea which was theoretically developed and empirically verified in multiple studies (most notably, Cox et al. (2007) and Cox et al. (2008)).

to take the form $u(m,G) = f(m)+g(G)$ (here f must be strictly increasing and g strictly decreasing under our monotonicity condition). What we have in mind is Debreu's "double cancellation condition" (Debreu, 1960) which in the current setting can be cast as follows: If $(m_1, G_1) \succeq (m_2, G_2)$ and $(m_2, G_3) \succeq (m_3, G_1)$, then $(m_1, G_3) \succeq (m_3, G_2)$. In words, the decision-maker's marginal disutility of guilt does not depend on how wealthy he is, and vice versa.[4]

Everything that has been said so far is either well-known or trivial in light of existing literature. In contrast, the next concept is new. Consider an agent who is indifferent between two options, one of which offers more money and more guilt (the greedy option), while the other offers less money and less guilt (the conscientious option). Formally, consider a pair of distinct alternatives $\{(m_1, G_1), (m_2, G_2)\}$ such that:

$$(m_1, G_1) \sim (m_2, G_2). \tag{4}$$

We call any $\{(m_1, G_1), (m_2, G_2)\}$ that satisfies (4) a *moral dilemma*. For a given moral dilemma, the agent is willing to give up $m_1 - m_2$ in order to reduce his level of guilt by $G_1 - G_2$. The sacrifice ratio between the two,

$$\frac{m_1 - m_2}{G_1 - G_2}, \tag{5}$$

is referred to as the *price of a clear conscience*. Our first moral axiom makes the (bold) postulate that the price of a clear conscience is independent of the moral dilemma the agent faces:

Axiom 1 (Constant Price of a Clear Conscience) *For any* $(m_1, G_1) \sim (m_2, G_2)$ *and* $(m_1', G_1') \sim (m_2', G_2')$*, it holds that* $\frac{m_1 - m_2}{G_1 - G_2} = \frac{m_1' - m_2'}{G_1' - G_2'}$.

The linear utility representation $u(m,G) = m - \theta G$ of equation (2) is easily shown to imply Axiom 1. In addition, the underlying preferences are clearly rational, monotone, and continuous. More surprisingly, the converse is true as well:

Theorem 1 *Consider a rational, monotone, and continuous preference relation* $\succeq$ *on* $\mathfrak{M} \times \mathfrak{G}$*. Then* $\succeq$ *admits the utility representation* $u(m,G) = m - \theta G$*,* $\theta > 0$ *if and only if* $\succeq$ *satisfies Axiom 1.*

Proof: In the Appendix.

Theorem 1 tells us that assuming a linear utility representation amounts to assuming that the mental tradeoff between money and guilt captured by the price of a clear conscience remains the same regardless of the level of income and load of sin an agent faces. This, it may be argued, is not necessarily a realistic description of actual behavior. We would arguably expect an agent to care less about giving up a unit of income the richer he is. He will therefore be willing to pay more to clear his conscience than a relatively poorer version of himself. The next axiom formalizes this description of behavior in a specific way by requiring the relative price of a clear conscience to always equal the relative income. This statement is of course only meaningful if relative income is well defined, *i.e.* if the increase in income is the same for both the greedy and the conscientious options.

[4]For further interpretation and discussion of this condition as well as equivalent independence type conditions see Debreu (1960) as well as Segal and Sobel (2002) and Vind and Grodal (2003).

Axiom 2 (Income Effect in Moral Choice) *For any* $(m_1, G_1) \sim (m_1', G_1')$ *and* $(m_2, G_2) \sim (m_2', G_2')$, *such that* $m_1/m_2 = m_1'/m_2'$, *it holds that* $\frac{m_1'-m_2'}{G_1'-G_2'} / \frac{m_1-m_2}{G_1-G_2} = m_1'/m_1$.

Axiom 2 implies the double cancellation condition. Indeed, in the Appendix we prove the following lemma:

Lemma 1 *Suppose a rational, monotone and continuous* $\succeq$ *on* $\mathfrak{M} \times \mathfrak{G}$ *satisfies Axiom 2. Then* $(m_1, G_1 + a) \sim (m_2, G_2 + a)$ *whenever* $(m_1, G_1) \sim (m_2, G_2)$, *and* $(m_1, G_1 + a) \succ (m_2, G_2 + a)$ *whenever* $(m_1, G_1) \succ (m_2, G_2)$ *for any* $a \in [\max\{-G_1, -G_2\}, +\infty]$.

Applying Lemma 1 to the antecedent statements of the double cancellation condition, we obtain:

$$(m_1, G_1) \succeq (m_2, G_2) \quad \Rightarrow \quad (m_1, G_3) \succeq (m_2, G_2 + G_3 - G_1) \tag{6}$$

$$(m_2, G_3) \succeq (m_3, G_1) \quad \Rightarrow \quad (m_2, G_3 + G_2 - G_1) \succeq (m_3, G_2) \tag{7}$$

which by transitivity yields $(m_1, G_3) \succeq (m_3, G_2)$. In particular, Axiom 2 implies that u must be additively separable. One can show that Axiom 1 also implies the double cancellation condition and therefore an additively separable utility function.

It turns out that Axiom 2 is necessary and sufficient for a representation which is logarithmic in money and linear in guilt.

Theorem 2 *Consider a rational, monotone and continuous preference relation* $\succeq$ *on* $\mathfrak{M} \times \mathfrak{G}$. *Then* $\succeq$ *admits the utility representation* $U(m, G) = \log m - \theta G$, $\theta > 0$ *if and only if* $\succeq$ *satisfies Axiom 2.*

Proof: In the Appendix.

Intuitively, the proof of Theorem 2 proceeds by showing that the indifference curves of $\succeq$ satisfying Axiom 2 are related by parallel displacement along the guilt axis and proportional expansion along the money axis, which leads to the desired representation.[5]

4 Reinterpreting Existing Literature

In this section, we continue the discussion of section 2 with the representation (3) in hand. We study three well-known laboratory games in which experimental subjects appear to be motivated by considerations other then their monetary payoff. Non-monetary motivation is manifested in positive giving in the Dictator game, non-zero contributions in the Public Good Provision game, and cooperation in the Prisoner's Dilemma. We show that this seemingly irrational behavior can be explained by guilt aversion, as modeled by functions (2) and (3). Indeed, both models admit observed experimental outcomes in equilibrium. However, we also demonstrate that the explanatory power of model (2) stems from its unfalsifiability, and prove that the logarithmic representation (3) achieves sharper predictions without sacrificing the goodness of fit.

[5] An alternative axiomatization of utility functions (2) and (3) is possible, where they rely on a common axiom requiring that the price of a clear conscience is constant for all moral dilemmas with fixed levels of money. Two additional axioms are then required to obtain Theorems 1 and 2 respectively, which describe how the PCC reacts to the change in money, holding the level of guilt fixed. This axiomatization can be found in a working paper version of this study.

The rest of this section proceeds as follows. First, for each game, we introduce the classic structure: Players, strategies and "material" payoffs corresponding to experimental payouts. Second, we replace these payoffs with utility functions (2) and (3), which depend not only on the strategies (via "material" payoff m), but also on pre-game beliefs (via guilt G). The extended utility function domain means that the resulting structure is a "psychological game" (Geanakoplos et al., 1989). If a player's preferences are represented by such a utility function, his preference ordering over the outcomes of the game, *e.g.* possible Dictator's donations to the Recipient, depends on his pre-game beliefs, *e.g.* what he thought the Respondent expected to receive. Third, the new game is solved for Psychological Nash equilibria, in which all players best-respond to their beliefs, and these beliefs are correct. Finally, the equilibria of the game under utility functions (2) and (3) are compared to experimental outcomes, which constitutes an empirical test of the models.

4.1 The Dictator Game

In the Dictator game, one of the players (the Dictator, hereafter D) determines how to split a total endowment T between himself and a passive player (the Recipient, hereafter R). D's donation to R, which is his strategy, is denoted by m_R. If D is selfish, *i.e.* if his preferences are reflected by his material payoff $T - m_R$, the only Nash equilibrium of the game is zero donation: $m_R = 0$. We would expect a guilt-averse D to donate a positive amount to R. This is formally confirmed below, when we find equilibria in psychological games induced by the linear (2) and the logarithmic (3) guilt models. Denote D's belief about R's expectation of the donation by $E(m_R)$. The amount of guilt that he experiences from donating m_R is thus $G = \max\{0, E(m_R) - m_R\}$, and Psychological Nash equilibrium (hereafter PsyNE) is a strategy m_R^* which solves (8).[6]

$$\begin{cases} U(m_R^*, E(m_R)) \geq U(m_R, E(m_R)) \text{ for all } m_R \in [0, T]; \\ m_R^* = E(m_R). \end{cases} \tag{8}$$

First, consider D's utility under the linear guilt model (2):

$$u = T - m_R - \theta \max\{0, E(m_R) - m_R\}. \tag{9}$$

PsyNE is determined from the system (10), which is obtained by applying the utility function (9) to the equilibrium condition (8):

$$\begin{cases} T - m_R^* - \theta \max\{0, E(m_R) - m_R^*\} \geq T - m_R - \theta \max\{0, E(m_R) - m_R\} \text{ for all } m_R \in [0, T], \\ m_R^* = E(m_R). \end{cases} \tag{10}$$

which yields

[6]Strictly speaking, Psychological Nash Equilibrium consists of a strategy profile and a belief profile (Geanakoplos et al., 1989). However, since these profiles are required to coincide (in equilibrium, strategies match beliefs), hereafter we denote PsyNE by its constituent strategy profile, implying that it is accompanied by the matching belief profile.

$$m_R^* \in \begin{cases} [0,T] & \text{if } \theta \geq 1; \\ \{0\} & \text{otherwise.} \end{cases} \tag{11}$$

Equation (11) characterizes the set of equilibria in the game. This set, depending on D's guilt sensitivity θ, either coincides with the strategy set $[0,T]$, or consists of the unique equilibrium where D gives zero.

Observation 1 *In the Dictator game with the linear utility function* (2)*, all possible Dictator's donations are PsyNE of the game if* $\theta \geq 1$*, and zero donation is the only PsyNE if* $\theta < 1$*.*

Figure 1, left illustrates the equilibria in the Dictator game under the linear guilt function, as solved from the system (10). The first equation in (10) solves for D's optimal donation m_R^* as a function of his belief about R's expectation $E(m_R)$. Depending on D's guilt sensitivity, it is either a 45° line from the origin $m_R^* = E(m_R)$ (if $\theta \geq 1$, solid graph), or a horizontal line $m_R^* = 0$ (if $\theta < 1$, dashed graph). The second equation in (10) is a 45° line from the origin, and the solution is their intersection, which is the whole line for $\theta \geq 1$ or a single point $(0,0)$ for $\theta < 1$.

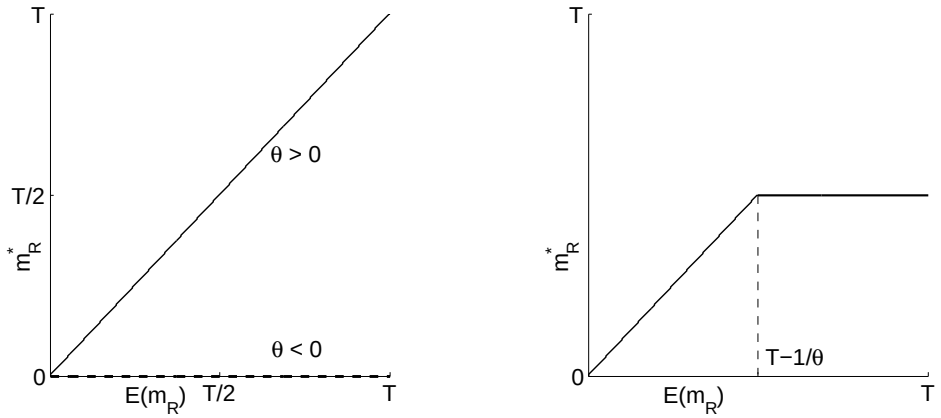

Figure 1: Dictator's optimal donation m_R^* as a function of his belief $E(m_R)$ under the linear (left) and logarithmic (right) guilt models.

Let us compare this prediction with the laboratory evidence. In a meta study of 129 different Dictator Game experiments (a total of 41,433 observations), Engel (2011) reports a mean contribution of 28.35% of the initial endowment. In fact, only 36.11% of all participants give nothing. Thus, if the linear guilt model (2) is correct, then for the majority of players guilt sensitivity must be larger than one. The model then suggests that these people will be giving out donations of all sizes, but the experimental data indicates that dictators are more likely to give little. Indeed, the distribution of average giving compiled in Engel (2011) is left skewed. Large donations are very rare. In particular, less than 10% of all participants surveyed in the meta-study gave more than 60% of the pie. In short, the linear model (2) does not account for the main stylized fact implied by the experimental evidence: The prevalence of moderate donations.

It turns out that the logarithmic model (3) embraces this stylized fact, predicting that for big expectations $E(m_R)$, D's optimal donation to R will be less than such expectation: $m_R^*(E(m_R)) < E(m_R)$.

It follows that such $E(m_R)$ cannot be part of an equilibrium and will not be observed. Indeed, PsyNE under model (3) solves the following system:

$$\begin{cases} \log(T - m_R^*) - \theta \max\{0, E(m_R) - m_R^*\} \geq \log(T - m_R) - \theta \max\{0, E(m_R) - m_R\} \text{ for all } m_R \in [0, T]; \\ m_R^* = E(m_R); \end{cases} \tag{12}$$

which yields

$$m_R^* \in [0, \max\{0, T - \frac{1}{\theta}\}]. \tag{13}$$

Expression (13) tells us that if D's preferences are represented by the logarithmic utility function (3), he will only satisfy expectations $E(m_R)$ up to a maximum of $T - \frac{1}{\theta}$ (Figure 1, right). This threshold expectation is increasing with D's guilt sensitivity.

Observation 2 *In the Dictator game with the logarithmic utility function*(3)*, the Dictator's donations up to* $\max\{0, T - \frac{1}{\theta}\}$ *are PsyNE of the game.*

Proof: In the Appendix.

Observation 2 implies that smaller donations are more likely to be observed, which is in line with the existing body of laboratory evidence. Note that, apart from the equilibrium prediction, the logarithmic model describes how D best-responds to a big $E(m_R)$: The amount he gives will be less than what he believes is expected from him. This implication of the logarithmic model validates the guilt aversion hypothesis, which has been put into question by some experimental results, most notably Ellingsen et al. (2010) who show that players' donations in the Dictator game do not always match their beliefs. Ellingsen et al. (2010) data is indeed inconsistent with the linear guilt model (2), which predicts that, for any belief $E(m_R)$, the Dictator will either grant it in full or give nothing. This leads Ellingsen et al. (2010) to refute the guilt aversion hypothesis. However, the logarithmic model, as we just argued, accounts for "sub-belief giving" by the Dictators, thus explaining most of their data.

4.2 The Public Good Provision Game

In a 2-player Public Good Provision game, each player $i = 1, 2$ is endowed with w_i and decides upon the amount of his contribution to a common fund $x_i \in [0, w_i]$, which is hence his strategy. The money in the fund is multiplied by a number $2a$ and shared equally among the players. The final payoff m_i is then determined as follows:

$$m_i = w_i - x_i + a(x_i + x_j), \tag{14}$$

where $1 > a > 0.5$. The restriction on a makes contributions collectively efficient but not individually rational.

If utilities equal material payoffs, the only equilibrium is zero contribution by both agents ("free-riding"). This sharp prediction is refuted by laboratory tests of the game, which report an average

contribution of 40-60% of the initial endowment (Ledyard, 1995). Can the phenomenon of positive contributions be attributed to guilt aversion? It has been shown that non-binding promises during pre-play communication helps sustain high contribution levels in Public Good Provision experiments (Ledyard (1995) and more recently Denant-Boemont et al. (2011)), which suggests the players' desire to meet expectations, *i.e.* guilt aversion. We formally confirm this intuition below, demonstrating that the models of guilt aversion (2) and (3) indeed admit positive contributions in equilibrium.

Let $E(x_i)$ denote i's second-order belief about x_i, *i.e.* what he thinks j expects him to contribute. First, consider the linear guilt model (2), under which i's utility becomes

$$\begin{aligned} u_i(x_i, x_j, E(x_i)) &= w_i - x_i + a(x_i + x_j) - \theta_i \cdot \max\{0, (w_j - x_j + a(E(x_i) + x_j)) - (w_j - x_j + a(x_i + x_j))\} \\ &= w_i - (1-a)x_i + ax_j - \theta_i \cdot \max\{0, a(E(x_i) - x_i)\}. \end{aligned}$$

Observe that utility is linear in the choice variable, which implies a corner solution. Indeed, it is easy to show that, if $\theta_i \geq \frac{1-a}{a}$, a player i will maximize his utility by contributing as much as expected from him ($x_i = E(x_i)$), regardless of j's contribution. In other words, any belief about his contribution is self-fulfilling, which means that the set of equilibria coincides with the set of strategy profiles. If the player is not guilt-averse enough ($\theta_i < \frac{1-a}{a}$), he will contribute zero.

Observation 3 *Consider a Public Good Provision game with the linear utility function* (2)*. A contribution profile* (x_1, x_2) *is a PsyNE if it satisfies the following: (i)* $x_i = 0$ *for any* $i = 1, 2$ *such that* $\theta_i < \frac{1-a}{a}$*; (ii)* $x_i \in [0, w_i]$ *for any* $i = 1, 2$ *such that* $\theta_i \geq \frac{1-a}{a}$*.*

Proof: In the Appendix.

The linear guilt model thus predicts that some agents will be giving out positive contributions, but remains agnostic about their size, their correlation with the opponent's contribution, and the effect of the parameters of the model on the amounts given. However, the existing body of experimental evidence from Public Good Provision games has some clearly identifiable patterns. In an early meta study, Ledyard (1995) observes that contributions positively depend on Marginal per Capita Return (a in our model). In a survey of post-1995 experimental literature, Chaudhuri (2011) emphasizes two stylized facts: First, heterogeneity of players in terms of social preferences, and second, the prevalence of conditional cooperators in the subject pool, whose contributions positively depend on the average contribution in the group (in a 2-player setting considered here, this is equivalent to dependence on the opponent's contribution). We will now show that the logarithmic guilt model (3) accounts for all three of these stylized facts.[7]

Under the logarithmic model (3), agent i's utility becomes:

$$U_i(x_i, x_j, E(x_i)) = \log(w_i - (1-a)x_i + ax_j) - \theta_i \cdot \max\{0, a(E(x_i) - x_i)\}. \tag{15}$$

[7]Heterogeneity of players is also implied by the linear model, which suggests that the players fall into one of the two groups, depending on their guilt sensitivity: Those who contribute nothing and those who can contribute anything. The logarithmic model provides a sharper prediction, since it implies a continuum of contribution behaviour, where the player's maximum contribution positively and continuously depends on his guilt sensitivity.

PsyNE is a pair of contributions (x_1^*, x_2^*), where each x_i satisfies the following conditions:

$$\begin{cases} U_i(x_i^*, x_j^*, E(x_i)) \geq U_i(x_i, x_j^*, E(x_i)) \text{ for all } x_i \in [0, w_i] \text{ for } i = 1, 2; \\ x_i^* = E(x_i) \text{ for } i = 1, 2. \end{cases} \tag{16}$$

Solving the system (16) yields Observation 4.

Observation 4 *Consider a Public Good Provision game with the logarithmic utility function* (3)*. A contribution profile* (x_1^*, x_2^*) *is a Psychological NE if, for* $i = 1, 2$*, either* $x_i^* = 0$ *or* $0 < x_i^* \leq (w_i + ax_j^*)/(1-a) - (\theta_i a)^{-1}$*.*

Proof: In the Appendix.

In words, for given values of players' guilt sensitivities, admitted in equilibrium are contributions up to a certain limit. The player's maximum contribution positively and continuously depends on the marginal return a, his guilt sensitivity θ_i and his opponent's contribution x_j, which is an exact match of the stylized experimental facts discussed above. The predictions of the model are illustrated in Figure 2 for the cases of big (left) and small (right) initial endowments: Shaded areas are Psychological Nash equilibria of the Public Good Provision game.

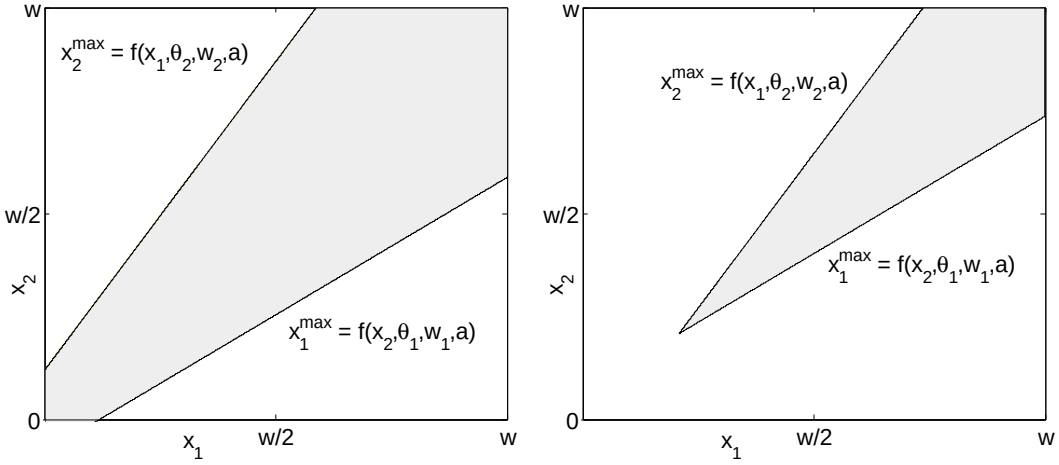

Figure 2: Equilibria in the Public Good Provision game under the logarithmic guilt model for the cases of big (left) and small (right) initial endowments, where $f(x_j, \theta_i, w_i, a) = (w_i + ax_j)/(1-a) - (\theta_i a)^{-1}$.

4.3 The Prisoner's Dilemma Game

In the Prisoner's Dilemma game, two players choose between cooperation and defection. The latter strategy is dominant but leads to an inefficient outcome, as described in the payoff table below:

	Cooperate	Defect
Cooperate	c, c	s, t
Defect	t, s	d, d

where $t > c > d > s$.[8]

Mutual defection is the only Nash equilibrium of Prisoner's Dilemma. Contrary to this prediction, existing experiments report non-negligible proportion of cooperative choices, around 20% in most studies (Andreoni and Miller, 1993; Cooper et al., 1996).

The presence of cooperation in laboratory Prisoner's Dilemma has been attributed to guilt aversion by some experimental papers, notably Miettinen and Suetens (2008). In Miettinen and Suetens (2008), it is argued that players who cooperate do so to avoid the feeling of guilt which comes with unilateral defection. We validate this intuition by applying the formal models of guilt aversion (2) and (3) to Prisoner's Dilemma and comparing their predictions with the experimental findings.

First, let us find pure strategy equilibria under the linear guilt model (2).

Observation 5 *Consider a Prisoner's Dilemma game with the linear utility function* (2)*. Mutual defection is a PsyNE for any values of* θ*; mutual cooperation is a PsyNE iff* $\theta_i \geq \frac{t-c}{c-s}$ *(for* $i = 1, 2$*) and unilateral cooperation, where* i *is the cooperator, is a PsyNE iff* $\theta_i \geq \frac{d-s}{t-d}$*.*

Proof: In the Appendix.

Observation 5 tells us that all strategy profiles can be equilibria, including, perhaps surprisingly, unilateral cooperation. Indeed, sufficiently guilt averse players (with $\theta_i \geq \frac{d-s}{t-d}$) will be willing to cooperate even when they think their opponent is going to defect. In order to see how extreme the guilt sensitivity thresholds are, we calculated its numerical value for payoff tables used in some of the most famous Prisoner's Dilemma experiments (Table 1).

Table 1: Guilt Sensitivity Thresholds under the Linear Model

Paper	Minimum Guilt Sensitivity	
	Unilateral Cooperation	Mutual Cooperation
Andreoni and Miller (1993)	0.5	0.71
Bereby-Meyer and Roth (2006)	0.7	0.7
Friedman and Oprea (2012)	0.29	0.8

For all considered studies, the condition for mutual cooperation is at least as restrictive as for unilateral cooperation. We illustrate this result in Figure 3, left, which shows that unilateral cooperation is an equilibrium under strictly larger set of parameter values than mutual cooperation, and hence is more likely to be observed (Zones B,C,D VS Zone C only).[9]

Contrary to this prediction, existing experimental research suggests that most players who cooperate in Prisoner's Dilemma do so conditionally on their opponent also cooperating. Cooper et al. (1996) do not identify any unconditional cooperators in their subject pool, and estimate the fraction of conditional cooperators to be 12.5-22%. Croson (2000) finds that at least 51% of subjects reciprocate their counterpart's expected action *i.e.* play a conditional strategy. Brosig (2002) reports that in a face-to-face experiment 90% of cooperators switch to defection when they think the opponent is going to defect.

[8]The letters stand for **t**emptation, **c**ooperation, **d**efection, **s**ucker.

[9]The numerical values in Figure 3 were calculated using Bereby-Meyer and Roth (2006) payoffs. For the payoff tables used in the other two studies (Andreoni and Miller, 1993; Friedman and Oprea, 2012), the prevalence of unilateral cooperation is even greater.

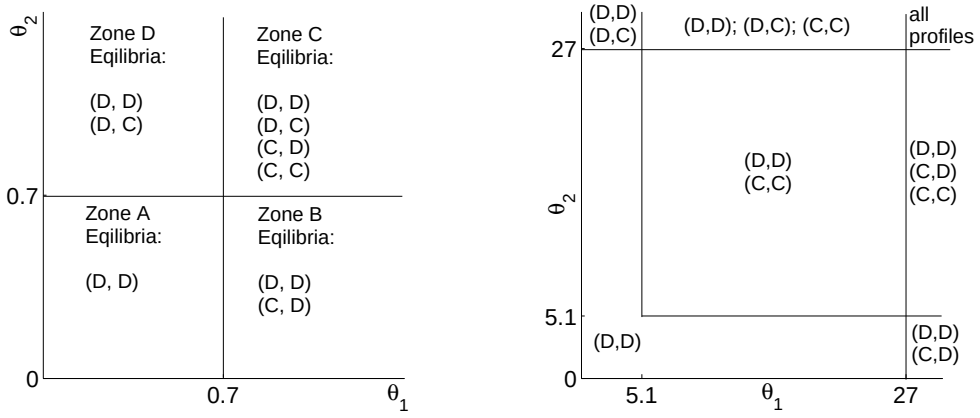

Figure 3: Equilibria in the Prisoner's Dilemma game under the linear (left) and logarithmic (right) guilt models.

Let us compare this evidence with the predictions of the logarithmic guilt model (3).

Observation 6 *Consider the Prisoner's Dilemma game with the logarithmic utility function* (3)*. Mutual defection is a PsyNE for any values of* θ*; mutual cooperation is a PsyNE when* $\theta_i \geq \frac{\log(t/c)}{c-s}$ *(for* $i = 1, 2$*) and unilateral cooperation, where* i *is a cooperator, is a PsyNE when* $\theta_i \geq \frac{\log(d/s)}{t-d}$*.*

Proof: In the Appendix.

Applying Observation 6 to experimental studies, we find that a necessary condition for mutual cooperation is strictly weaker than for unilateral cooperation (Table 2).[10] The infinity signs mean that a player will never be a unilateral cooperator in Andreoni and Miller (1993) and Friedman and Oprea (2012).[11] This is because under model (3) the tradeoff between money and guilt depends on the agent's wealth. A unilaterally cooperating player is poorer than a mutually cooperating one, hence he cares less about the harm he inflicts on the opponent by switching to defection. Consequently, unilateral cooperation is hard to sustain in equilibrium.

Table 2: Guilt Sensitivity Thresholds under the Logarithmic Model

Paper	Minimum Guilt Sensitivity	
	Unilateral Cooperation	Mutual Cooperation
Andreoni and Miller (1993)	∞	0.08
Bereby-Meyer and Roth (2006)	27	5.1
Friedman and Oprea (2012)	∞	0.06

Figure 3, right illustrates Observation 6. It shows that, even for Bereby-Meyer and Roth (2006) payoffs, where all strategy profiles can still be equilibria, unconditional cooperation is limited to the extreme end of guilt sensitivity distribution. As a result, (C,C) is a more likely equilibrium outcome than (D,C) and (C,D) combined; which is in line with the experimental evidence.

[10]Note that although the cardinal values of θ are not comparable between the linear and the logarithmic models, the order of the thresholds is meaningful

[11]This is because in these papers "sucker" payoff is equal to 0, which entails infinitely negative utility for the player under the logarithmic model. Hence, there does not exist an amount of guilt large enough to dissuade him from switching to defection in this case.

5 Conclusion and Discussion

By developing a formal theory of moral choice, we provided axiomatic foundations for two utility representations of guilt aversion. First, we proved a representation theorem for a frequently used linear utility, which enabled us to carefully examine and question its underlying axioms. We then proposed a logarithmic representation grounded in more realistic assumptions about the way agents trade off material and moral considerations. By applying the logarithmic representation to laboratory games, we showed that it can better account for the existing body of experimental evidence than a linear representation. Moreover, the novel functional form is able to address experimentally informed criticism of a guilt aversion hypothesis, *e.g.* it predicts Ellingsen et al. (2010) observation that actions do not always match beliefs. Our framework gives experimentalists tools to pick the most appropriate utility model by considering the axioms of moral choice. Future theoretical research can draw on our approach to discover the grounding of other utility models of non-selfish preferences.

Our analysis can be extended to dynamic frameworks; in particular, income received today can affect guilt experienced tomorrow. Guilt in round n can be modeled as a payoff loss inflicted on the other player in round $n-1$. In fact, the lasting nature of emotions is well-known in psychological literature (Davidson, 1998; Gratz et al., 2010) and was also observed in laboratory repeated games (Grimm and Mengel, 2011; Madarász et al., 2012; Neo et al., 2013). Coupled with our non-linear utility, this lag in guilt can explain consumption smoothing, as well as end-effects in experimental games (e.g. "cutoff" cooperation in repeated Prisoner's Dilemma).

One potential problem with the current paper, as well as with the established models of guilt aversion, is that any utility function over "monetary" payoffs has an implication for the agent's risk preferences.[12] In games with belief-dependent utility, these implications are poorly understood. It seems reasonable to conjecture that models linear in money (Battigalli and Dufwenberg, 2007; López-Pérez, 2010; Miettinen, 2013) imply risk neutrality of the agent, whereas our logarithmic specification corresponds to risk aversion. In this paper, however, we remain agnostic about the exact risk properties of our model, owing to our focus on pure strategy equilibria. The implicit risk assumptions of guilt averse preferences deserve further investigation in a separate study.

Appendix

Representation Theorems

In order to establish Theorems 1 and 2, we prove several auxiliary results, which allow us to construct new indifference sets from existing ones. This follows from continuity of $\succeq$ and Axioms 1-2, which describe how given indifference sets are related to one another.

We first establish Lemmas 1-4 which are used in the proof of Theorem 2.

Our first lemma demonstrates that the indifference sets of $\succeq$ satisfying Axiom 2 are related by parallel displacements along the guilt axis.

[12] We are grateful for an anonymous referee for pointing this out.

Proof of Lemma 1. First, observe that Axiom 2 implies that the PCC is constant for moral dilemmas with fixed amounts of money: $\frac{m_2-m_1}{G_2-G_1} = \frac{m_2-m_1}{G_2'-G_1'}$ for any $(m_1, G_1) \sim (m_2, G_2)$ and $(m_1, G_1') \sim (m_2, G_2')$.

Second, we prove the symmetric ($\sim$) part of the lemma. Consider $(m_1, G_1) \sim (m_2, G_2)$. WLOG, let $m_2 > m_1$ (then $G_2 > G_1$ by monotonicity). Suppose $(m_1, G_1 + a) \not\sim (m_2, G_2 + a)$. First, suppose $(m_1, G_1 + a) \succ (m_2, G_2 + a)$. By monotonicity, $(m_2, G_2 + a) \succ (m_1, G_2 + a)$. Since by continuity the lower and the upper contour sets of $(m_2, G_2 + a)$ are closed, their intersections with a closed connected interval $\{(m_1, G) : G \in [G_1 + a, G_2 + a]\}$ are also closed. Moreover, they are non-empty, since the ends of the interval belong to the upper and the lower contour sets respectively, as shown above. Recall that a connected space cannot be divided into two disjoint non-empty closed sets, hence there is a point (m_1, G_3), with $G_1+a < G_3 < G_2+a$ (*) which belongs to both intersections, *i.e.* $(m_1, G_3) \sim (m_2, G_2+a)$. Hence we have two moral dilemmas which, by Step 1, should entail the same price of clear conscience: $\frac{m_2-m_1}{G_2-G_1} = \frac{m_2-m_1}{G_2+a-G_3}$ which implies $G_3 = G_1 + a$, a contradiction to (*). Now suppose that $(m_1, G_1 + a) \prec (m_2, G_2 + a)$. There are two cases to consider. If $a > 0$, then $(m_2, G_2 + a) \prec (m_2, G_2) \sim (m_1, G_1)$. Thus, $(m_1, G_1) \succ (m_2, G_2 + a) \succ (m_1, G_1 + a)$. By connectedness and continuity again, there exists a point (m_1, G_4) such that $(m_2, G_2 + a) \sim (m_1, G_4)$, where $G_1 + a > G_4 > G_1$ (**). Then by Step 1, $G_2 - G_1 = G_2 + a - G_4$, *i.e.* $G_4 = G_1 + a$, which contradicts (**). The case when $a < 0$ is demonstrated analogously.

Third, we prove the asymmetric ($\succ$) part of the lemma. Let $(m_1, G_1) \succ (m_2, G_2)$. We need to show that $(m_1, G_1+a) \succ (m_2, G_2+a)$. *Case 1.* $G_1 > G_2$ (hence $m_1 > m_2$) and $a > 0$. Suppose, conversely, that $(m_1, G_1+a) \preceq (m_2, G_2+a)$. By monotonicity and transitivity, $(m_1, G_1+a) \preceq (m_2, G_2+a) \prec (m_2, G_2) \prec (m_1, G_1)$. By connectedness of $\{(m_1, G) : G \in [G_1, G_1 + a]\}$ and continuity of $\succeq$, there exists such G_3 that $(m_2, G_2 + a) \sim (m_1, G_3)$ and $G_1 < G_3 \leq G_1 + a$ (*). Now consider (m_2, G_2). By connectedness and continuity again, there exists G_4 such that $(m_2, G_2) \sim (m_1, G_4)$ and $G_1 < G_4 < G_1 + a$. By Step 1, $G_4 - G_2 = G_3 - (G_2 + a)$ and thus $G_3 = G_4 + a > G_1 + a$, but by (*) $G_3 \leq G_1 + a$, which is a contradiction. Thus, our supposition was wrong, and $(m_1, G_1 + a) \succ (m_2, G_2 + a)$. The remaining three cases are demonstrated analogously.

In order to be able to construct new indifference sets from existing ones by proportional expansion along the money axis, we need to prove existence of an indifference curve passing through any two money levels, which is demonstrated in the following lemma.

Lemma 2 *If a rational, monotone and continuous $\succeq$ satisfies Axiom 2, then for any (m, G) and any $m' > m$, there exists G' such that $(m, G) \sim (m', G')$.*

Proof. The proof relies on continuity of $\succeq$, connectedness of $\mathfrak{M}$ and quasilinearity in guilt (established in Lemma 1), and is available on request.

We can now use Lemma 2 to show that indifference sets of $\succeq$ satisfying Axiom 2 are related by proportional expansion along the money axis. This is established in Lemma 3.

Lemma 3 *If a rational, monotone and continuous $\succeq$ on $\mathfrak{M} \times \mathfrak{G}$ satisfies Axiom 2, then $(bm_1, G_1) \sim (bm_2, G_2)$ whenever $(m_1, G_1) \sim (m_2, G_2)$ and $(bm_1, G_1) \succ (bm_2, G_2)$ whenever $(m_1, G_1) \succ (m_2, G_2)$ for any $b \in R_{++}$.*

Proof. First, we prove the symmetric part of the lemma. Consider $(m_1, G_1) \sim (m_2, G_2)$. WLOG, let $m_1 > m_2$. By Lemma 2, $\exists G_3$ so that $(bm_2, G_2) \sim (bm_1, G_3)$. By Axiom 2, $\frac{bm_1 - bm_2}{G_3 - G_2} \Big/ \frac{m_1 - m_2}{G_1 - G_2} = b$, which implies that $G_3 = G_1$. Hence, $(bm_1, G_1) \sim (bm_2, G_2)$.

Second, we prove the asymmetric part of the lemma. Let $(m_1, G_1) \succ (m_2, G_2)$. *Case 1:* $m_1 > m_2$. By Lemma 2, $\exists G_3$ and G_4 so that $(m_2, G_2) \sim (m_1, G_3)$ and $(bm_2, G_2) \sim (bm_1, G_4)$. By monotonicity, $G_1 < G_3$. By Axiom 2, $\frac{bm_1 - bm_2}{G_4 - G_2} \Big/ \frac{m_1 - m_2}{G_3 - G_2} = b$, which implies that $G_4 = G_3$. Hence, by monotonicity, $(bm_2, G_2) \sim (bm_1, G_3) \prec (bm_1, G_1)$. The remaining case $(m_2 > m_1)$ is demonstrated analogously.

In the proof of Theorem 2, we will reduce the comparison of points on the money-guilt plane to the comparison of points on the m axis. In order to show it is possible, in Lemma 4 we prove that every point in $\mathfrak{M} \times \mathfrak{G}$ belongs to an indifference set which has a nonempty intersection with any line parallel to the m axis (including the m axis itself).

Lemma 4 *If a rational, monotone and continuous $\succeq$ satisfies Axiom 2, then for any (m, G) and any G', there exists m' such that $(m, G) \sim (m', G')$.*

Proof. The proof is similar to that of Lemma 2 with two divergences. First, the roles of the two variables (m and G) are switched. Second, we are able to demonstrate the existence of m' for two cases: $G' > G$ and $G' < G$. In the proof we require Lemma 3 in the same way as Lemma 1 was required to prove Lemma 2.

Finally, we establish two auxiliary results necessary to prove Theorem 1. The first of these shows that indifference curves of $\succeq$ satisfying Axiom 1 are related by parallel displacement along the money axis.

Lemma 5 *If a rational, monotone and continuous $\succeq$ on $\mathfrak{M} \times \mathfrak{G}$ satisfies Axiom 1, then $(m_1 + a, G_1) \succ (m_2 + a, G_2)$ whenever $(m_1, G_1) \succ (m_2, G_2)$ for any $a \in [\max\{-m_1, -m_2\}, +\infty]$.*

Proof. Similarly to the asymmetric part of Lemma 1, this proof consists of 4 cases. However, in the rest of the Appendix we will only be using the result for one case: $m_2 > m_1$ and $a > 0$, which is considered here. The remaining 3 cases are proved analogously. Let $(m_1, G_1) \succ (m_2, G_2)$, where $m_1 < m_2$ (hence $G_1 < G_2$ by monotonicity). We need to show that $(m_1 + a, G_1) \succ (m_2 + a, G_2)$, where $a > 0$. Suppose, conversely, that $(m_1 + a, G_1) \preceq (m_2 + a, G_2)$. By monotonicity, $(m_1 + a, G_1) \succ (m_1, G_1)$ and hence by transitivity $(m_1 + a, G_1) \succ (m_2, G_2)$. By connectedness of $\{(m, G_2) : m \in [m_2, m_2 + a]\}$ and continuity of $\succeq$, there exists such m_3 that $(m_1 + a, G_1) \sim (m_3, G_2)$ and $m_2 < m_3 \leq m_2 + a$ (*). Now consider (m_1, G_1) and observe that $(m_2, G_2) \prec (m_1, G_1) \prec (m_2 + a, G_2)$, where the last relation follows from monotonicity and transitivity. By connectedness and continuity again, there exists m_4 such that $(m_1, G_1) \sim (m_4, G_2)$ and $m_2 < m_4 < m_2 + a$. By Axiom 1, $m_4 - m_1 = m_3 - (m_1 + a)$ and thus $m_3 = m_4 + a > m_2 + a$, but by (*) $m_3 \leq m_2 + a$, which is a contradiction. Thus, our supposition was wrong, and $(m_1 + a, G_1) \succ (m_2 + a, G_2)$, q.e.d.

In the proof of Theorem 1, we will require a result similar to Lemma 2 for preferences satisfying Axiom 1, *i.e.* that for any initial money-and-guilt situation (m, G) and any large guilt $G' > G_1$ there

exist a sum of money large enough to "seduce" the agent by making him indifferent between the status quo and the large guilt.

Lemma 6 *If a rational, monotone and continuous preference relation $\succeq$ on $\mathfrak{M} \times \mathfrak{G}$ satisfies Axiom 1, then for any $(m, G) \in \mathfrak{M} \times \mathfrak{G}$ and any $G' > G$, there exists $m' \in \mathfrak{M}$ such that $(m, G) \sim (m', G')$.*

Proof. The proof is similar that of Lemma 2, with the exception that the roles of the variables (m and G) are switched. In the proof we require Lemma 5 in the same way as Lemma 1 was required to prove Lemma 2.

We are now ready to establish our main representation theorems.

Proof of Theorem 1. Necessity is established in the text before the theorem. We now show sufficiency. Suppose $\succeq$ satisfies Axiom 1. Consider the point $(1, 0)$. By Lemma 6, there exists $m^* \in \mathfrak{M}$ such that $(1, 0) \sim (m^*, 1)$. By monotonicity of $\succeq$, such m^* is unique. Let $\theta = \frac{m^*-1}{1-0} = m^* - 1$. Assign $u(m, G) = m - \theta G$ uniquely for all $(m, G) \in \mathfrak{M} \times \mathfrak{G}$. We now show that u represents $\succeq$, *i.e.* that $u(m_1, G_1) \geq u(m_2, G_2)$ iff $(m_1, G_1) \succeq (m_2, G_2)$. First, suppose $(m_1, G_1) \succeq (m_2, G_2)$. Pick a $G_3 > \max\{G_1, G_2\}$. By Lemma 6, there exist μ_1 and μ_2 such that $(m_1, G_1) \sim (\mu_1, G_3)$ and $(m_2, G_2) \sim (\mu_2, G_3)$. By Axiom 1, $\frac{\mu_1-m_1}{G_3-G_1} = \frac{\mu_2-m_2}{G_3-G_2} = \theta$. Then $u(m_1, G_1) = m_1 - \theta G_1 = \mu_1 - \theta(G_3 - G_1) - \theta G_1 = \mu_1 - \theta G_3 \geq \mu_2 - \theta G_3 = \mu_2 - \theta(G_3 - G_2) - \theta G_2 = m_2 - \theta G_2 = u(m_2, G_2)$. Second, suppose $u(m_1, G_1) \geq u(m_2, G_2)$ and consider μ_1, μ_2 defined as above. Then $\mu_1 = m_1 + \theta(G_3 - G_1) = u(m_1, G_1) + \theta G_3 \geq u(m_2, G_2) + \theta G_3 = m_2 + \theta(G_3 - G_2) = \mu_2$. By monotonicity it follows that $(\mu_1, G_3) \succeq (\mu_2, G_3)$ and then by transitivity $(m_1, G_1) \succeq (m_2, G_2)$, q.e.d.

Proof of Theorem 2. *Necessity.* Suppose $U(m, G) = \log m - \theta G$ represents $\succeq$ for some $\theta > 0$. Consider $(m_1, G_1) \sim (m_2, G_2)$ and $(m_1', G_1') \sim (m_2', G_2')$, where $m_1/m_1' = m_2/m_2'$. Then $\log m_1 - \theta G_1 = \log m_2 - \theta G_2$ and $\log m_1' - \theta G_1' = \log m_2' - \theta G_2'$, which implies $\theta(G_1 - G_2) = \log \frac{m_1}{m_2} = \log \frac{m_1'}{m_2'} = \theta(G_1' - G_2')$. The relative PCC then becomes $\frac{m_1-m_2}{G_1-G_2} \Big/ \frac{m_1'-m_2'}{G_1'-G_2'} = \frac{m_1-m_2}{(m_1-m_2)m_1'/m_1} = \frac{m_1}{m_1'}$, which establishes Axiom 2.

Sufficiency. Step 1. Consider any $(m_1, G_1) \in \mathfrak{M} \times \mathfrak{G}$. If $G_1 = 0$, let $U(m_1, G_1) = \log m_1$, which is uniquely defined. If $G_1 > 0$, by Lemma 4, $\exists \mu_1 \in \mathfrak{M}$ s.t. $(m_1, G_1) \sim (\mu_1, 0)$. Moreover, such $\mu_1(m_1, G_1)$ is unique. Indeed, suppose $(m_1, G_1) \sim (\mu_1, 0)$ and $(m_1, G_1) \sim (\mu_1', 0)$, where $\mu_1 \neq \mu_1'$. WLOG, let $\mu_1 > \mu_1'$. Then, by monotonicity, $(\mu_1, 0) \succ (\mu_1', 0)$, which contradicts transitivity of $\succeq$. Assign $U(m_1, G_1) = \log \mu_1(m_1, G_1)$.

We now show that U defined as above represents $\succeq$. Suppose $(m_2, G_2) \succeq (m_1, G_1)$. As shown above, there exist unique numbers μ_1 and μ_2 such that $(\mu_2, 0) \sim (m_2, G_2) \succeq (m_1, G_1) \sim (\mu_1, 0) \Rightarrow \mu_2 \geq \mu_1 \Rightarrow U(m_2, G_2) = \log \mu_2 \geq \log \mu_1 = U(m_1, G_1)$ by transitivity and monotonicity. Following the same steps backwards proves that $(m_2, G_2) \succeq (m_1, G_1)$ whenever $U(m_2, G_2) \geq U(m_1, G_1)$.

It is left to be demonstrated that U is of the specified form, *i.e.* there exists a well-defined function $f(G) = U(m, G) - \log m$, $\forall G$. In order to do so, consider (m_1, G_1) and (m_3, G_3) such that $G_1 = G_3$. We need to show that $f(G_1) = f(G_3)$. WLOG, let $m_1 > m_3$. By Lemma 4, there exist μ_1 and μ_3 such that $(m_3, G_3) \sim (\mu_3, 0)$ and $(m_1, G_1) \sim (\mu_1, 0)$. By Lemma 3, $(m_3, G_3) \sim (\mu_3, 0)$ implies

$(m_1, G_1) \sim (\frac{\mu_3 m_1}{m_3}, 0)$, thus $\mu_1 = \frac{m_1}{m_3}\mu_3$. Thus, $f(G_1) = U(m_1, G_1) - \log m_1 = \log \mu_1 - \log m_1 = \log \frac{\mu_1}{m_1} = \log \frac{m_1 \mu_3}{m_3 m_1} = \log \mu_3 - \log m_3 = U(m_3, G_3) - \log m_3 = f(G_3)$.

Step 2. We need to demonstrate that f defined as above is linear in G. It suffices to show that, for any G_1, G_2, G_3 and $G_4 \in \mathfrak{G}$ such that $G_2 - G_1 = G_4 - G_3$, it is true that $f(G_2) - f(G_1) = f(G_4) - f(G_3)$.

Consider arbitrary G_1, G_2, G_3 and m, as well as $G_4 = G_3 + G_2 - G_1$. WLOG, let $G_2 > G_1$ and $G_3 > G_1$. As shown in Step 1, there exist unique μ_1 and μ_2 such that $(m, G_1) \sim (\mu_1, 0)$ and $(m, G_2) \sim (\mu_2, 0)$. Thus, $f(G_1) = \log \mu_1 - \log m$ and $f(G_2) = \log \mu_2 - \log m$, which implies $f(G_2) - f(G_1) = \log \frac{\mu_2}{\mu_1}$. Lemma 4 ensures that there exists m^* such that $(m, G_1) \sim (m^*, G_3)$. Thus, by transitivity, $(m^*, G_3) \sim (\mu_1, 0)$, *i.e.* $U(m^*, G_3) = \log \mu_1$ and $f(G_3) = \log \mu_1 - \log m^*$.

By Lemma 1, $(m, G_1) \sim (m^*, G_3) \;\Rightarrow\; (m, G_2) \sim (m^*, G_3 + (G_2 - G_1))$, where $G_3 + (G_2 - G_1) = G_4$. We can now calculate $U(m^*, G_4) = U(m, G_2) = \log \mu_2$ and $f(G_4) = \log \mu_2 - \log m^*$. It is now easy to see that an arbitrary change in G produces the same change in f at any two values of G: $f(G_4) - f(G_3) = (\log \mu_2 - \log m^*) - (\log \mu_1 - \log m^*) = \log \frac{\mu_2}{\mu_1} = f(G_2) - f(G_1)$.

Thus, f can be written as $f = a + bG$. Observe that $a = f(0) = \log m - \log m = 0$. In order to calculate b, consider a point $(1, 1)$. Note that $f(1) = \log \mu_0 - \log 1 = \log \mu_0$, where $(1, 1) \sim (\mu_0, 0)$, and μ_0 is unique for a given preference relation. Hence $b = \mu_0$. We can now write $f(G) = \mu_0 G$. Observe that, by monotonicity, $\mu_0 < 1$ and $\log \mu_0 < 0$. Assign $\theta = -b = -\log \mu_0 > 0$.

$U(m, G) = \log m - \theta G$ represents $\succeq$, q.e.d.

Applications

Proof of Observation 2. Under the logarithmic guilt model (3), D's utility becomes

$$U(m_R, E(m_R)) = \log(T - m_R) - \theta \max\{0, E(m_R) - m_R\}. \tag{17}$$

D maximizes his utility by choosing an optimal donation m_R^*, given his second-order belief (*i.e.* his belief about R's expectation of the donation) $E(m_R)$. If the optimal donation equals the belief, it is a PsyNE of the game.

Let $m_R^*(E(m_R))$ denote D's optimal donation, as a function of his belief.

Observe that, due to the maximum operator present in the function form, $U(m_R)$ admits a kink at $m_R = E(m_R)$, where $U(m_R)$ is continuous, but not differentiable.

Note that the kink point $m_R = E(m_R)$ separates the function into two halves:

$$U = \begin{cases} \log(T - m_R) - \theta(E(m_R) - m_R) & \text{if } m_R \leq E(m_R); \\ \log(T - m_R) & \text{if } m_R > E(m_R). \end{cases} \tag{18}$$

Let $U_l(m_R) = \log(T - m_R) - \theta(E(m_R) - m_R)$ and $U_r(m_R) = \log(T - m_R)$.

Observe that $\frac{\partial U_l}{\partial m_R} = \frac{-1}{T - m_R} + \theta$. Denote by m_R^l the value of m_R maximizing the function U_l:

$$m_R^l = \begin{cases} T - \frac{1}{\theta} & \text{if } T - \frac{1}{\theta} > 0; \\ 0 & \text{if } T - \frac{1}{\theta} \leq 0. \end{cases} \tag{19}$$

Observe that utility to the right of the kink is decreasing: $\frac{\partial U_r}{\partial m_R} = \frac{-1}{T-m_R} < 0$. Also, utility to the left of the kink is concave: $\frac{\partial^2 U_l}{\partial m_R^2} = \frac{-1}{(T-m_R)^2} < 0$. The value of m_R maximizing the whole function is thus the smaller of the two: the kink $E(m_R)$, or the value m_R^l maximizing the left-hand-side utility U_l:

$$m_R^*(E(m_R)) = \begin{cases} 0 & \text{if } T - \frac{1}{\theta} \leq 0; \\ T - \frac{1}{\theta} & \text{if } 0 < T - \frac{1}{\theta} < E(m_R); \\ E(m_R) & \text{otherwise,} \end{cases} \tag{20}$$

which proves Observation 2.

Proof of Observation 4. Recall the equilibrium condition on contributions and beliefs in the Public Good Provision game:

$$x_i^* = \begin{cases} U_i(x_i^*, x_j^*, E(x_i)) \geq U_i(x_i, x_j^*, E(x_i)) \text{ for all } x_i \in [0, w_i] \text{ for } i = 1, 2; \\ E(x_i) \text{ for } i = 1, 2. \end{cases} \tag{21}$$

Similarly to the Dictator game application, the function $U_i(x_i)$ admits a kink at $x_i = E(x_i)$; it is strictly decreasing to the right of it ($\frac{\partial U_r}{\partial x_i} = \frac{\partial \log(w_i - (1-a)x_i + ax_j)}{\partial x_i} = \frac{a-1}{w_i-(1-a)x_i+ax_j} < 0$) and strictly concave to the left of it ($\frac{\partial^2 U_l}{\partial x_i^2} = \frac{\partial^2(\log(w_i-(1-a)x_i+ax_j)-\theta_i a(E(x_i)-x_i))}{\partial x_i^2} = \frac{-(a-1)^2}{(w_i-(1-a)x_i+ax_j)^2} < 0$). Hence, best response contribution is equal to the expectation if and only if either of the two conditions holds: (i) utility is increasing to the left of the kink: $\frac{\partial U_l}{\partial x_i}(E(x_i)) \geq 0$ or (ii) expectation is zero $E(x_i) = 0$. Below we solve for equilibrium contribution x_i^* under (i).

$$\begin{cases} \frac{\partial(\log(w_i-(1-a)x_i+ax_j)-\theta_i a(E(x_i)-x_i))}{\partial x_i}(E(x_i)) \geq 0; \\ x_i^* = E(x_i). \end{cases} \tag{22}$$

$$\begin{cases} (a-1)/(w_i - (1-a)E(x_i) + ax_j) + \theta_i a \geq 0; \\ x_i^* = E(x_i). \end{cases} \tag{23}$$

$$\begin{cases} x_i^* \leq (w_i + ax_j)/(1-a) - (\theta_i a)^{-1}; \\ x_i^* = E(x_i). \end{cases} \tag{24}$$

As one can see from the formula, player i's maximum equilibrium contribution is increasing with his initial endowment w_i, his guilt sensitivity θ_i and the other player's contribution x_j. It is also easy to show that x_i^* is increasing with return on investment a.

Proof of Observation 5. We will establish the values of parameters θ_i, $i = 1, 2$ under which each pure strategy profile (with corresponding beliefs) is a Psychological Nash equilibrium.

1. (*Defect, Defect*; $E(m_1) = d$, $E(m_2) = d$) is the only NE of the game and hence also an equilibrium in the psychological game with guilt aversion (deviations from it will neither increase the players' material payoffs nor decrease their guilt, which is already zero).

2. (*Cooperate, Defect*; $E(m_1) = s$, $E(m_2) = t$) is an equilibrium if:

$$\begin{cases} u_1(C,D) & \geq & u_1(D,D); \\ u_2(C,D) & \geq & u_2(C,C). \end{cases} \tag{25}$$

Applying utility function (2) and solving for θ_1, θ_2 we obtain

$$\begin{cases} \theta_1 & \geq & \frac{d-s}{t-d}; \\ \theta_2 & \in & R. \end{cases} \tag{26}$$

3. (*Cooperate, Cooperate*; $E(m_1) = c$, $E(m_2) = c$) is an equilibrium if $u_i(C,C) \geq u_i(D,C)$ for $i = 1, 2$. Applying utility function (2) we obtain $c - \theta_i(c-c) \geq t - \theta_i(c-s)$, which yields $\theta_i \geq \frac{t-c}{c-s}$.

Proof of Observation 6. Similarly to the previous proof, we write down conditions ensuring that unilateral deviation is unprofitable for each candidate pure strategies equilibrium.

1. (*Defect, Defect*; $E(m_1) = d$, $E(m_2) = d$) is always an equilibrium in games with guilt aversion as argued above.

2. (*Cooperate, Defect*; $E(m_1) = s$, $E(m_2) = t$) is an equilibrium if

$$\begin{cases} \log s - \theta_1(t-t) & \geq & \log d - \theta_1(t-d); \\ \log t - \theta_2(s-s) & \geq & \log c - \theta_2 \cdot 0; \end{cases} \tag{27}$$

which yields

$$\begin{cases} \theta_1 & \geq & \frac{\log(d/s)}{t-d}; \\ \theta_2 & \in & R. \end{cases} \tag{28}$$

2. (*Cooperate, Cooperate*; $E(m_1) = c$, $E(m_2) = c$) is an equilibrium if $\log c - \theta_i(c-c) \geq \log t - \theta_i(c-s)$ for $i = 1, 2$, which yields $\theta_i \geq \frac{\log(t/c)}{c-s}$.

References

Ahrens, S. and D. Snower (2014). Envy, guilt, and the Phillips curve. *Journal of Economic Behavior & Organization 99*, 69–84.

Andreoni, J. and J. Miller (1993). Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence. *Economic Journal 103*(418), 570–85.

Andreoni, J. and J. Rao (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics 95*(7), 513–520.

Attanasi, G. and R. Nagel (2007). A survey of psychological games: Theoretical findings and experimental evidence. *Games, Rationality and Behaviour. Essays on Behavioural Game Theory and Experiments, Palgrave McMillan, Houndmills*, 204–232.

Battigalli, P., G. Charness, and M. Dufwenberg (2013). Deception: The role of guilt. *Journal of Economic Behavior and Organization 93*, 227–232.

Battigalli, P. and M. Dufwenberg (2007). Guilt in Games. *American Economic Review 97*(2), 170–176.

Battigalli, P. and M. Dufwenberg (2009). Dynamic Psychological Games. *Journal of Economic Theory 144*(1), 1–35.

Bereby-Meyer, Y. and A. Roth (2006). The Speed of Learning in Noisy Games: Partiall Reinforcement and the Sustainability of Cooperation. *American Economic Review 96*(4), 1029–42.

Brosig, J. (2002). Identifying cooperative behavior: some experimental results in a prisoners dilemma game. *Journal of Economic Behavior & Organization 47*(3), 275–290.

Chang, L., A. Smith, M. Dufwenberg, and A. Sanfey (2011). Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion. *Neuron 70*(3), 560–572.

Charness, G. and M. Dufwenberg (2006). Promises and Partnership. *Econometrica 74*(6), 1579–1601.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics 14*(1), 47–83.

Cooper, R., D. DeJong, R. Forsythe, and T. Ross (1996). Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games. *Games and Economic Behavior 12*(2), 187–218.

Cox, J. C., D. Friedman, and S. Gjerstad (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior 59*(1), 17–45.

Cox, J. C., D. Friedman, and V. Sadiraj (2008). Revealed altruism. *Econometrica 76*(1), 31–69.

Croson, R. T. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior & Organization 41*(3), 299–314.

Davidson, R. J. (1998). Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & Emotion 12*(3), 307–330.

Debreu, G. (1954). *Representation of a Preference Ordering by a Numerical Function.* New York: Wiley.

Debreu, G. (1960). *Topological methods in cardinal utility.* Stanford, California: Stanford University Press.

Denant-Boemont, L., D. Masclet, and C. Noussair (2011). Announcement, observation and honesty in the voluntary contributions game. *Pacific Economic Review 16*(2), 207–228.

Ellingsen, T., M. Johannesson, S. Tjotta, and Torsvik (2010). Testing guilt aversion. *Games and Economic Behavior 68*(1), 95–107.

Engel, C. (2011). Dictator games: a meta study. *Experimental Economics 14*(4), 583–610.

Friedman, D. and R. Oprea (2012). A continuous dilemma. *American Economic Review*, 337–363.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological Games and Sequential Rationality. *Games and Economic Behavior 1*(1), 60–79.

Geng, H., A. Weiss, and I. Wolff (2011). The Limited Power of Voting to Limit Power. *Journal of Public Economic Theory 13*(5), 695–719.

Gratz, K. L., M. Z. Rosenthal, M. T. Tull, C. Lejuez, and J. G. Gunderson (2010). An experimental investigation of emotional reactivity and delayed emotional recovery in borderline personality disorder: The role of shame. *Comprehensive Psychiatry 51*(3), 275–285.

Grimm, V. and F. Mengel (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters 111*(2), 113–115.

Güth, W., M. Ploner, and T. Regner (2009). Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology 30*(5), 814–827.

Hopfensitz, A. and E. Reuben (2009). The Importance of Emotions for the Effectiveness of Social Punishment. *Economic Journal 119*(540), 1534–1559.

Ketelaar, T. and W. Au (2003). The Effects of Feelings of Guilt on the Behavior of Uncooperative Individuals in Repeated Social Bargaining Games: An Effect-as-information Interpretation of the Role of Emotion in Social Interaction. *Cognition and Emotion 17*(17), 429–453.

Khalmetski, K. (2015). Testing guilt aversion with an exogenous shift in beliefs. *Available at SSRN 2363163*.

Khalmetski, K., A. Ockenfels, and P. Werner (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*.

Ledyard, J. (1995). Public goods: some experimental results. In J. Kagel and A. Roth (Eds.), *Handbook of Experimental Economics*. Princeton: Princeton University Press.

López-Pérez, R. (2010). Guilt and shame: an axiomatic analysis. *Theory and Decision 69*(4), 569–586.

Madarász, K., U. Gneezy, and A. Imas (2012). Conscience accounting: emotional dynamics and social behaviour. Available at: `http://eprints.lse.ac.uk/47994/`. Accessed: 2016-01-31.

Miettinen, T. (2013). Promises and conventions–an approach to pre-play agreements. *Games and Economic Behavior 80*, 68–84.

Miettinen, T. and S. Suetens (2008). Communication and Guilt in a Prisoner's Dilemma. *Journal of Conflict Resolution 52*(6), 945–960.

Neilson, W. (2006). Axiomatic Reference-Dependence in Behavior toward Others and toward Risk. *Economic Theory 28*(3), 681–692.

Neo, W. S., M. Yu, R. A. Weber, and C. Gonzalez (2013). The effects of time delay in reciprocity games. *Journal of Economic Psychology 34*, 20–35.

Sandbu, M. (2008). Axiomatic foundations for fairness-motivated preferences. *Social Choice and Welfare 31*(4), 589–619.

Segal, U. and J. Sobel (2002). Min, max, and sum. *Journal of Economic Theory 106*(1), 126–150.

Vind, K. and B. Grodal (2003). *Independence, Additivity, Uncertainty*. Berlin: Springer.